



(12) **EUROPEAN PATENT APPLICATION**

(43) Date of publication:  
**03.09.2003 Bulletin 2003/36**

(51) Int Cl.<sup>7</sup>: **G06F 17/30**

(21) Application number: **03250903.6**

(22) Date of filing: **13.02.2003**

(84) Designated Contracting States:  
**AT BE BG CH CY CZ DE DK EE ES FI FR GB GR**  
**HU IE IT LI LU MC NL PT SE SI SK TR**  
Designated Extension States:  
**AL LT LV MK RO**

(30) Priority: **26.02.2002 US 82354**

(71) Applicants:  
• **Jacobsen, Brian**  
**2720 Vanlose (DK)**  
• **Soren, Alain**  
**1623 Copenhagen W. (DK)**

(72) Inventors:  
• **Jacobsen, Brian**  
**2720 Vanlose (DK)**  
• **Soren, Alain**  
**1623 Copenhagen W. (DK)**

(74) Representative: **Lucas, Brian Ronald**  
**Lucas & Co.**  
**135 Westhall Road**  
**Warlingham, Surrey CR6 9HJ (GB)**

(54) **Subject specific search engine**

(57) A subject-specific search engine utilizes a smart web crawler and includes a capability of filtering out sites not relevant to the particular subject. As the smart crawler traverses the Internet, sites are filtered, and only sites found relevant are indexed and stored in a database for later searching. Sites may be filtered an

arbitrary number of times for relevance, and such filtering may, for example, comprise automated, lexicon-based filtering; manual filtering, using a human editor; or both

## Description

**[0001]** The present invention is directed to the general field of systems, methods, and computer program products for performing Internet-based searching. In particular, it deals with a search engine tailored to search the Internet and to return results that contain fewer irrelevant results than present search engines return.

**[0002]** It has been said that the Internet/network communities are what are pushing the economy forward these days, and it is a fact, that the Internet contains unprecedented volumes of information on just about any topic. The only problem is to find the truly relevant resources. Search engines are what make the Internet useful, because without these tools the chances of finding relevant resources would be significantly diminished. Thus, while the Internet drives the economy, search engines drive the Internet. This is backed by statistics made on users' use of the Internet, which shows that users spend more online time at search engines than anywhere else, including portals.

**[0003]** Yet current search engine technology often leaves one dissatisfied and frustrated, particularly where one would like to find resources on a given subject in a specific context. For example, suppose that a user would like to find information on the Ford Pinto in a legal context (referring to the product liability cases against Ford due to defects in the Ford Pinto models design). A general purpose search engine (GPSE) will typically return numerous irrelevant links if one searches on the term "Pinto," simply because a GPSE can not recognize a context or a specific subject, e.g. a legal context or law as a subject. This is so due to the fact that GPSEs adopt the strategy of "everything is relevant;" therefore, they try to collect and index all pages on the Internet. Their operations are based on this unedited collection of pages.

**[0004]** To gain more insight into the workings of GPSEs, it is first worth noting that the term "search engine" is typically used to cover several different types of search facilities. In particular, "search engines" may be broken up into four main categories: robots/crawlers; metacrawlers; catalogs with search facilities; and catalogs or link collections.

**[0005]** Figure 1A illustrates the operation of robots/crawlers. These are characterized by having a process (i.e., a crawler) that traverses the Internet 1, as indicated by arrow 4, on a site-by-site basis and sends back to its host machine 2 the contents of each home page it encounters at various sites 3 on its way, as indicated by arrows 5. Then, as shown in Figure 1B, the host machine 2 indexes the pages 8 sent back by crawler 7 and files the information in its database 9. Any front-end query looks up the search terms in the information stored in the host's database 9. Existing crawlers generally consider all information to be relevant, and therefore, all home pages on all sites traversed are indexed. Examples of such robots/crawlers include Google™, Altavis-

ta™, and Hotbot™.

**[0006]** Metacrawlers, as illustrated in Figure 2, are characterized in that they offer the possibility of searching in a single search facility 2 and obtaining replies from multiple search facilities 10. The metacrawler serves as a front end to several other facilities 10 and does not have its own "back end." Metacrawlers are limited by the quality of the information in the search facilities that they employ. Examples of such metacrawlers include MetaCrawler™, LawCrawler™, and LawRunner™.

**[0007]** Catalogs, with or without search facilities, are characterized in that they are collections of links structured and organized by hand. In the case of a catalog with a search facility, a front end query results in the system looking up the search terms in the information manually stored in the host machine's database. In the case of a catalog without a search facility, a user must gather information manually by clicking through the links in the catalog and assessing whether or not they are relevant. A catalog is limited by the capacity of its editors and by their priorities. Examples of catalogs include Yahoo!™, Jubii™ (the Danish equivalent of Yahoo!™), and FindLaw™.

**[0008]** Catalogs fall under the category of "portals" and "vortals." Portals and, to some extent, proprietary databases like FindLaw.com™ and WestLaw.com™ try to solve the problem of finding relevant information in different ways. Portals try to provide an overview of selected sites manually, by letting editors "surf the Internet" and gather links to relevant resources/sites. These editors are able to scan /view and evaluate 10-25 sites a day, only one or two of which will typically be of the desired quality. This approach is ineffective if one's goal is to provide an overview of the Internet, more so since portals will usually only provide links to sites' start/main pages.

**[0009]** Vertical portals (vortals), which are portals concentrating on particular subject matter, have all the same problems, only more acutely, because they must be more precise in their qualification and labeling processes. This makes the task of reaching a critical mass of contents and references even harder and more time consuming. An example of such a vortal is FindLaw.com™, which has evolved since its startup in 1995.

**[0010]** Most GPSEs consist of a crawler, an indexing part, and a front-end for queries. A typical GPSE works in the following way. A GPSE has a database of links to home pages. The crawler picks a link, downloads the page, and saves it in memory. It then takes the next link, downloads the page and on and on. The indexing part then reads one of the saved pages from memory and analyzes its content. If there are links on the page, they are saved in the database so that the crawler can fetch those pages later. How the content of the page itself is indexed depends on the particular GPSE. A user can enter a query into the front-end, and the GPSE will search the indexed pages. This procedure is based on the principle of "everything is relevant," meaning that the

crawler will get and save every page it encounters. Similarly every page saved in memory by the crawler will be indexed. This typical operation of a GPSE is illustrated in Figures 1A and 1B, as discussed above (indexing part not shown).

**[0011]** The present engine takes the form of a subject specific search engine (SSSE), where the strategy adopted is to collect and index only the pages deemed relevant for a specific subject, e.g., law or medicine. The way this is done, in one embodiment of the invention, is through a lexicographic analysis of the texts used by the profession or area of interest. The inventive technology is able to differentiate among contexts and to thus provide a given profession with a search engine that returns only links to relevant resources, i.e., resources the contents of which contain a query term in the desired context. Drawing from the "Pinto" example discussed above, a search for "Pinto" in a legal search engine according to the present invention will thus only return results where the query term "Pinto" appears in a legal context. Put another way, it will return only legal documents or legally relevant documents containing the term "Pinto".

**[0012]** To further understand the advantages of an SSSE according to the invention over GPSEs, consider the following scenario:

**[0013]** Imagine a public library that keeps all of its books in a huge pile. Suppose an attorney needs to find some information about the product liability case brought against Ford for their design faults in the Pinto model.

**[0014]** Now imagine that the attorney goes to the public library. In this public library all the books are placed in the back, and in order to retrieve any books one must approach a librarian and tell her which word one is looking for. In this case the attorney is looking for "Pinto." In less than a second the librarian is back and places 460,000 books in front of the attorney. Depending on the public library these books may not be ordered at all, ordered by the number of times the word "Pinto" appears in each book, or by other people's references to each book.

**[0015]** To find a book covering the Pinto case, suppose that the attorney starts looking at the title of each book, and if it seems interesting he reads the back cover. It will not be long before he finds himself reading about Pinto horses, families having Pinto as a surname, the El Pinto restaurant, etc. Once in a while he will find a book that actually is about the Ford Pinto case. If he has the patience and the time, he will find the type of book he is looking for somewhere along the line. If he is able to scan through the books at a rate of one book per second, he will be finished in approximately five and a half days. The end result may be some 500 books.

**[0016]** To avoid this, the attorney currently has two choices:

- He can use Boolean algebra, if he is familiar with it,

by changing the query to something like "'Ford Pinto" AND ("product liability" OR "punitive damages").' To ensure that he gets all the relevant books, he should also enter all kinds of legal terms (the U.S. legal terminology consists of approximately 20,000 terms).

- He can find the "legal librarian" (or, in Internet terms, a metacrawler, like LawRunner.com or LawCrawler.com). The legal librarian does some of the work that the attorney must do in the preceding option, that is, the librarian makes sure that both the original word, "Pinto," and either the word "legal" or the word "law" is in the books that the librarian returns. It might, however, seem a bit inadequate to get only two terms out of the 20,000 terms mentioned above (i.e., without entering the rest manually).

**[0017]** The attorney may, however, have a third option, a specialized library (for example, a university department's library, like a law school library or an engineering school library). A specialized library is a library specializing in one subject; in the present example, the appropriate library would be a law library. If the attorney were to ask the librarian here, the "Pinto" query would result in, perhaps, 500 books. The key here is that, before any book is placed in the specialized library, it has been classified as relevant in the library's particular context. That is, someone actually sat down, looked through the book, and decided that it contained relevant material. As a result, in the present example, all the books about Pinto horses, families, and the like, never make it into this library, thus eliminating the hassle of ignoring them.

**[0018]** The inventive SSSE draws upon some of the concepts of this third option. In particular, the inventive SSSE provides a particular profession (or more generally, special interest group) with a search engine that returns only links to resources that contain components of the profession's terminology.

**[0019]** The inventive SSSE starts with the principle that not all pages or even sites are relevant. If one is building an SSSE for United States law, pages from sites like [www.games.com](http://www.games.com) and [www.mp3.com](http://www.mp3.com) are generally not relevant. A human would "know" that a site with the name [www.games.com](http://www.games.com) is not likely to contain pages with a relevant content for the legal profession. The question then is how to make a computer system "know."

**[0020]** In an SSSE according to an embodiment of the present invention, a first feature is that the crawler may perform filtering and indexing, in addition to merely finding information. This means that the crawler is now "aware" of the analysis of each web page and can act accordingly.

**[0021]** A second feature of an SSSE according to an embodiment of the present invention is the addition of a new field in the database containing the information stored by the crawler. This field holds a parameter re-

ferred to as the "depth." The depth is the number of preceding pages that were traversed and were deemed not relevant.

**[0022]** A third feature of an SSSE according to an embodiment of the present invention is the setting of a threshold for how deep the crawler will be permitted to crawl down a branch before it is stopped. That is, how many irrelevant pages in a row will be allowed before the branch may be considered entirely irrelevant.

**[0023]** In one embodiment of the invention, the crawler is designed so as to filter each site it traverses using a database of relevant terminology. In another embodiment of the invention, the information is sent to the host, and all analyzing processes are left to the host computer running the crawler. The web page corresponding to each site that is passed through the filter and deemed preliminarily relevant may then be filtered one or more additional times. Filtering may be performed either automatically or in conjunction with a human. In the case of automatic filtering, the additional filtering may be performed either as part of the crawler or as a process running on a host computer. Pages that are passed through as many filtering stages as are present and are deemed relevant are then indexed and stored in a database.

**[0024]** To provide users with ease in retrieving the most relevant information, an embodiment of the invention utilizes a ranking system for determining which pages are most relevant. The ranking system is based on the computation and storage of word rankings and the computation of site (page) rankings, based on the word rankings, in response to user queries. Rankings are then used to display the sites retrieved in the search in accordance with their rankings, so as to give display priority to the most relevant sites.

**[0025]** Also for the sake of user-friendliness, an embodiment of the invention utilizes a hierarchical display system. For example, all pages linked to from a given page may be displayed indented under the main page's URL. Such a display may be implemented in collapsible/expandable form. As discussed above, display may take into account site rankings.

**[0026]** The invention may be embodied in the form of a method, system, and computer program product (i.e., on a computer-readable medium).

#### Definitions of Terms

**[0027]** In describing the invention, the following definitions are applicable throughout (including above).

**[0028]** A "computer" refers to any apparatus that is capable of accepting a structured input, processing the structured input according to prescribed rules, and producing results of the processing as output. Examples of a computer include a computer; a general-purpose computer; a supercomputer; a mainframe; a super mini-computer; a mini-computer; a workstation; a microcomputer; a server; an interactive television; a hybrid combination of a computer and an interactive television; and

application-specific hardware to emulate a computer and/or software. A computer can have a single processor or multiple processors, which can operate in parallel and/or not in parallel. A computer also refers to two or more computers connected together via a network for transmitting or receiving information between the computers. An example of such a computer includes a distributed computer system for processing information via computers linked by a network.

**[0029]** A "computer-readable medium" refers to any storage device used for storing data accessible by a computer. Examples of a computer-readable medium include a magnetic hard disk; a floppy disk; an optical disk, like a CD-ROM or a DVD; a magnetic tape; a memory chip; and a carrier wave used to carry computer-readable electronic data, such as those used in transmitting and receiving e-mail or in accessing a network.

**[0030]** "Memory" refers to any medium used for storing data accessible by a computer. Examples include all the examples listed above under the definition of "computer-readable medium."

**[0031]** "Software" refers to prescribed rules to operate a computer. Examples of software include software; code segments; instructions; computer programs; and programmed logic.

**[0032]** A "computer system" refers to a system having a computer, where the computer comprises a computer-readable medium embodying software to operate the computer.

**[0033]** A "network" refers to a number of computers and associated devices that are connected by communication facilities. A network involves permanent connections such as cables or temporary connections such as those made through telephone or other communication links, or both. Examples of a network include an internet, such as the Internet; an intranet; a local area network (LAN); a wide area network (WAN); and a combination of networks, such as an internet and an intranet.

#### BRIEF DESCRIPTION OF THE DRAWINGS

**[0034]** Embodiments of the invention will now be described with reference to the attached drawings in which:

Figures 1A and 1B together illustrate the operation of a typical prior-art GPSE, and Figures 1A also partially illustrates the operation of a crawler according to an embodiment of the present invention;

Figure 2 illustrates the operation of a typical prior-art metacrawler;

Figures 3A and 3B illustrate, along with Figure 1A, the operation of an embodiment of an SSSE according to the present invention;

Figure 4 illustrates a configuration according to an embodiment of the present invention;

Figures 5A and 5B illustrate a depth monitoring process according to an embodiment of the inven-

tion;

Figures 6A, 6B, and 6C illustrate various embodiments of filtering operations according to the present invention;

Figure 7 illustrates an exemplary process used in implementing steps of the embodiments shown in Figures 6A, 6B, and 6C; and

Figures 8A and 8B depict exemplary display formats according to embodiments of the invention.

## DESCRIPTION OF EMBODIMENTS OF THE INVENTION

**[0035]** The general structure of an embodiment of an SSSE according to the invention is shown in Figure 4. As shown, there are three primary components in the SSSE: smart crawler 16, host computer 17, and human interface 18. Smart crawler 16 operates, for the most part, as shown in Figure 1A (that is, similar to prior-art crawler programs); however, there are additional features that differentiate the inventive smart crawler from prior-art crawlers discussed above. As is the case with typical GPSEs, as discussed above, the crawler, in this case smart crawler 16, transmits information back to host computer 17; this is similar to host machine 2 in Figure 1A, but it may also perform additional processes. Finally, human interface 18 is provided for entering search queries and for, in some embodiments, human interaction in the processes of information screening and indexing. The roles of these components will become clearer in view of the discussion below explaining the operation of the inventive SSSE.

**[0036]** As explained above, in one embodiment smart crawler 16 operates in basically the same way as prior-art crawlers, i.e., by visiting sites and transmitting information back to host computer 17. However, unlike prior-art crawlers, in this embodiment smart crawler 16 does not operate under the "everything is relevant" principle; rather, it operates as shown in Figure 3A. In Figure 3A, smart crawler 16 traverses the Internet 1 and then performs a screening operation, denoted by site filter 11. Site filter 11 determines, based on terminology of the profession to which the SSSE is directed (e.g., law), whether or not each site is considered to be relevant to the profession. The result is that some sites are filtered out, leaving, essentially, an Internet 1' containing only relevant sites. It is only the information on relevant sites that is transmitted to host computer 17 in this embodiment. At host computer 17, the information on relevant sites, as determined by site filter 11, may be stored in memory (not shown) for further processing, or it may be indexed and stored in a database. Filter 11 may be implemented in either automated form or in a form requiring human interaction.

**[0037]** In another embodiment the filtering capabilities may be implemented solely in the host computer 17. In this case, smart crawler 16 returns all site information to host computer 17 for screening, and host computer

17 makes all determinations as to whether or not sites are relevant and as to when links to sites should be traversed or not. As in the case of the previous embodiment, filtering may be automated or manual (e.g., human editing).

**[0038]** Figure 3B reflects further steps that may be carried out in some embodiments of the process carried out by the inventive SSSE. In such embodiments, there is at least one additional level of filtering 13, which may be carried out either as part of smart crawler 16 or as a separate process in host computer 17. As shown, the information (web page) 12 of each site found relevant in the process shown in Figure 3A may be screened at least a second time, by filter 13, which again screens the terminology found in the site information (in the case of a manual implementation, a human may also be able to account for additional site information, like the name of the site and the "overall feel" of the site). Only information 14 that passes through this second filter 13 is then indexed 15 and stored in database 9.

**[0039]** Therefore, in general, each site stored in the SSSE passes through one or more layers of screening, each of which may be implemented in automated or manual form. In one exemplary embodiment, two automated filters are followed by human screening prior to indexing.

**[0040]** As discussed above, filter 13 acts to filter out irrelevant web pages (and similarly with any additional filters, if present). The pages that are filtered out are discarded, and no links to such pages are traversed. Thereafter, if the smart crawler encounters a link to a discarded page, it simply ignores it.

The strategy of using both a smart crawler having automated filtering and human editing in some embodiments of the invention combines the best of two worlds: the speed of the machine and the reasoning of man. The machine suggests a number of sites, the editor approves or discards the sites, and the machine indexes the relevant pages on the approved sites. Based on links from the approved sites, the smart crawler may suggest more sites, etc., resulting in an evolution of the search engine.

**[0041]** A further difference between an embodiment of the inventive smart crawler and prior-art crawlers lies in the use of a "depth monitoring system" in connection with determining whether or not sites should even be visited. Figures 5A and 5B will be used to describe such a depth monitoring system, according to an embodiment of the invention.

**[0042]** Figure 5A depicts a "chunk" of the Internet and will be used as an aid in explaining Figure 5B. Figure 5A depicts a hierarchy of three levels: i, j, and k. Each level has at least one site. Additionally, links between sites will be referred to below as "link<sub>xy</sub>," where "xy" designates the fact that the link is from site x to site y. Note that the Internet, when viewed on a larger scale, is generally not a hierarchy; however, on a small scale, as depicted, it can be viewed as such. In any event, the in-

ventive system is applicable to the Internet, in general.

**[0043]** Figure 5B gives a flowchart demonstrating the operation of the inventive depth monitoring system. Assume that site *i* has already been visited and that a link from site *i* to a site in the *j*<sup>th</sup> level, say, *j*<sub>1</sub>, has been traversed (i.e., link<sub>*ij*1</sub> has been traversed). When a site is initially visited, each of its links to further sites is assigned a depth equal to that of the link that was traversed to reach the site, i.e.,  $D_{link_{jk}} = D_{link_{ji}}$  for each outgoing link (*jk*; for this example, links *j*<sub>1</sub>*k*<sub>1</sub> and *j*<sub>1</sub>*k*<sub>2</sub>); this is shown in step S1. In step S2, filter 11 makes a determination as to whether or not the site visited (in this example, *j*<sub>1</sub>) is relevant. If so, the process goes to step S7, where the depths of all outgoing links from the site are reset to zero (i.e., in the ongoing example, step S7 would set  $D_{link_{j1k1}} = 0$  and  $D_{link_{j1k2}} = 0$ ). The process then traverses a link to the next level S8 (e.g., from *j*<sub>1</sub> to *k*<sub>1</sub>); in so doing, the current site (e.g., *j*<sub>1</sub>) will next be considered to be the previous site (that is, *i* becomes *j*<sub>1</sub>), and the next site (e.g., *k*<sub>1</sub>) will be considered to be the current site (that is, *j* becomes *k*<sub>1</sub>). From here, the process returns to step S1. If step S2 determines that the current site (e.g., *j*<sub>1</sub>) is not relevant, then step S3 increments  $D_{link_{jk}}$  for all outgoing links (in the example, *j*<sub>1</sub>*k*<sub>1</sub> and *j*<sub>1</sub>*k*<sub>2</sub>). The process then proceeds to step S4, where it is determined whether or not  $D_{link_{jk}}$  for the outgoing links from the site exceeds a predetermined maximum value,  $D_{max}$ . If  $D_{link_{jk}} > D_{max}$ , then no sites stemming from that site are visited, and the links from that site are deleted S5; that is, the "branch" ends at that site. If this were the case, then the next site at the same level of the hierarchy (here, *j*<sub>2</sub>) would be visited (if there were no such site, then the process would go back to the previous level to determine if there were another site to be visited from there, etc.) S6. In general, depth is monitored for each link traversed, until it is determined that at least one link from the original site leads to at least one relevant site (i.e., within a depth of no more than  $D_{max}$ ; if this never happens, then all links from the site are deleted).

**[0044]** To understand this process more fully, consider the following additional example, where  $D_{max}$  is assumed to be two:

1. A page A contains a link to the page B. Page A is deemed relevant, so the link to B has depth 0.
2. B has a link to C. B is deemed not relevant, so the link to C is assigned the depth 1.
3. C has a link to D. C is deemed not relevant, so the link to page D is assigned the depth 2.
4. D has a link to E. D is deemed relevant, so the link to page E is assigned the depth 0.
5. Note that if D had been deemed not relevant, the link to page E would have been assigned the depth 3, which is greater than  $D_{max}$ . In this case, the link from D to E would have been deleted, and it would have been determined if there were another site to be visited from C.

**[0045]** In a more concrete example, suppose there is a link to [www.games.com](http://www.games.com) and that the SSSE is geared toward the legal context. It is most likely that [www.games.com](http://www.games.com) would be deemed not relevant in a legal context, so all the links from [www.games.com](http://www.games.com) to other pages, both on [www.games.com](http://www.games.com) and other sites, would have the depth 1. Suppose further that from [www.games.com](http://www.games.com), the smart crawler follows the link to

[www.games.com/Review\\_The\\_](http://www.games.com/Review_The_Ultimate_Car_Game.html)

[Ultimate\\_Car\\_Game.html](http://www.games.com/Review_The_Ultimate_Car_Game.html),

which has a link to

[www.joysticks.com](http://www.joysticks.com),

from which there are further links. The link from

[www.games.com](http://www.games.com)

to

[www.games.com/](http://www.games.com/Review_The_Ultimate_Car_Game.html)

[Review\\_The\\_Ultimate\\_Car\\_Game.html](http://www.games.com/Review_The_Ultimate_Car_Game.html)

will be given a depth of 1, and the link from this page to

[www.joysticks.com](http://www.joysticks.com)

will be given a depth of 2. If the maximum depth is set to 2, and if the page [www.joysticks.com](http://www.joysticks.com) is deemed not relevant, the links from [www.joysticks.com](http://www.joysticks.com) are discarded (again assuming a maximum depth of 2).

**[0046]** The embodiment of the invention discussed above makes use of at least one automated filter. Exemplary embodiments of automated filtering are depicted in Figures 6A-6C. Figure 6A shows the basic idea of the exemplary embodiment of automated filtering according to the invention. A web page is input to the filter, and there is an optional step S11 of removing extraneous material, that is, of recognizing and eliminating from consideration things like advertisements. The main part of the filter takes the page and compares it with a lexicon of terms (e.g., legal terms) S12 whose presence will indicate that the page may be relevant. If the comparison is favorable (this will be discussed further below) S13, then the page is saved S15. If not, then the page is discarded S14.

**[0047]** Figure 6B shows a second exemplary embodiment of automated filtering. In this embodiment, prior to any analysis, the page under consideration is broken up into component parts (cells) S16. This serves the purpose of making it easier to discriminate between material that needs to be tested and material that is extraneous S11. It also permits a piecemeal approach to testing. The components are passed into a test stage, where first it is determined if there are any remaining components that need to be tested S17. If yes, then the next component is compared with the lexicon S12', and the process loops back to S17. If not, then all components of the page have been tested, and the question is asked as to whether or not there was at least one relevant component on the page S18. If not, then the page is discarded S14. If yes, then the page is saved S15.

**[0048]** Figure 6C depicts a fully component-oriented exemplary embodiment of automated filtering. As in Figure 6B, the web page is broken up into its constituent components S16, and extraneous components may be

removed S11. The process then determines if there is still a component of the page left to test S17. If not, the process ends S19. Otherwise, the next component is compared with the lexicon S12'. The process then determines if the comparison results are favorable S20. If yes, then the component is saved S22; if not, then it is discarded S21. In this manner, the database that is built by the SSSE need only perform queries on relevant portions of pages, rather than on entire pages that may include irrelevant material.

**[0049]** The above embodiments all include steps of comparing with a lexicon and determining whether or not the comparison was favorable. Figure 7 depicts an exemplary embodiment of how this may be done. For each object (page or component) to be tested, each word, term, or expression in the object is compared with the words, terms, and expressions found in the lexicon S20. Within the lexicon, different words, terms, and expressions may be assigned different weights, for example, according to relative significance. If it is determined that a word, term, or expression in the object matches one in the lexicon, the weight assigned to the word, term, or expression is added to a cumulative total weight for the object S21. Once the entire object has been tested in this fashion, the cumulative total weight is compared to a predetermined threshold value S22. The value of the threshold may be set according to how selective the SSSE designer wants the database to be. If the cumulative total weight exceeds the threshold, the object is deemed relevant and is saved S24. If not, then the object is deemed irrelevant and is discarded S23.

**[0050]** Also in two of the above embodiments is the step of breaking up a web page into components S16. In an exemplary embodiment, this may be done by splitting up each web page into cells, where a cell is a portion of the page. This is done by analyzing the HTML code for the page. In one embodiment, cells may correspond to paragraphs of text; however, they may correspond to any desired components of the web page (e.g., lines of text or different portions of a page having multiple areas of text). In addition to the advantages of breaking up a web page into its components S11 discussed above, this also makes it easier to remove extraneous material, like menus, banners, etc., leaving only the cells containing material that might contain relevant text.

**[0051]** One particular advantage to using a lexicon-based filter is that all of the components of the filter may be the same for any context/profession, except for the lexicon. Therefore, one need only change the lexicon accessed by the other components in order to create a search engine for a different context/profession. This may be done within the host computer 17 (in Figure 4) by referencing a different memory for each context/profession. This may, in turn, be done by referencing a different file in a memory (for example, on a hard drive of the host computer) or by replacing a replaceable memory component (for example, a floppy disk or a CD-ROM).

**[0052]** In one embodiment of the invention, an inventive site ranking feature is also included. This ranking system analyzes the Internet (i.e., the sites found) to determine the degree to which sites have been found interesting by others in the desired context/profession. In particular, in one embodiment, this is determined by finding the number of links and citations to sites from other relevant sites determined by a user query. This information may be used in conjunction with displaying the results of the query, in order to emphasize the sites most likely to be helpful.

**[0053]** In a further embodiment of the invention, the site ranking feature is implemented using a word ranking scheme. The basic idea of this technique is to assign numerical scores to words and to sum the scores of the words on a page to determine a score for the page. The technique works by examining each word (non-trivial word, i.e., not "stop words," like "if," "it," "and," and the like) on a given page and increasing its score if it appeared on a relevant page (i.e., a page that passed through filtering) containing a link to the given page. In a sub-embodiment, the word score is increased according to how many relevant pages that linked to the given page contain the word. In a further embodiment, the technique is augmented by increasing a word's score according to where it appears in a link leading to the page being examined. In particular, if the word appears closer in proximity to the link to the page being examined, its score is increased.

**[0054]** A word score is saved for each word on each page (i.e., except for stop words, as discussed above). When a user enters a query, the inventive SSSE determines a set of (relevant) pages that contain the query terms. For each page, the word scores are summed for the words of the query to compute a site ranking for that page. The site rankings for the pages are then used in determining how to display the search results to the user. In summary, the inventive system utilizes dynamic site rankings, computed based on word rankings and in response to user queries.

**[0055]** A further feature according to an embodiment of the invention is a user-friendly display of results. In a preferred embodiment, this user-friendly display is a hierarchical type of display. In a further embodiment, the display uses the site ranking feature to determine an order in which to display the results. That is, the most relevant sites, as determined by their rankings, would be displayed earlier in the display and/or more prominently than less high ranking sites.

**[0056]** Figures 8A and 8B show two exemplary embodiments of a display according to the present invention. Figure 8A shows a display in the form of a file-document type of hierarchy. A file 20, 22 represents a type of pages/sites that it contains. As shown, the type may contain additional sub-types (shown as files). The type also contains documents, which represent the actual pages/sites. One traverses the hierarchy by clicking on files 20, 22 to open them until one locates a desired doc-

ument 21. One then clicks on the document 21 to access the information or site.

**[0057]** Similarly, Figure 8B shows a display in menu form. In the depiction of Figure 8B, there are six "site types" that represent six different classes of information found during a search of the SSSE database. As in conventional menu-based system, if there is an arrow in the menu, that indicates another level of menu. In the example shown, a user has clicked on Site Type F to reveal three sites (F1, F2, and F3). The user may then access any particular one of these sites by clicking on the appropriate menu item.

**[0058]** In a further embodiment of the invention, the "files" or "site types" in one level of the hierarchy may consist of URLs of sites, and the next level of the hierarchy may then contain "files"/ "site types" and "documents"/ "sites" linked to from those URLs.

**[0059]** Note that, while the invention has been described above in the context of the Internet, it may be similarly applied to any other computer network.

**[0060]** The inventive procedure is based on the principle that "most pages are not relevant" and that the inventive SSSE should separate the "straw" from the "chaff." This permits the inventive system not to visit every page on the Internet because it can quickly determine that a site is not relevant and, as a result, all the pages on that site are not indexed. One of the consequences of this is that highly irrelevant pages, like most "free home pages," are discarded. Another consequence is that the inventive system builds a fairly large database of relevant material very rapidly.

**[0061]** The invention has been described in detail with respect to preferred embodiments, and it will now be apparent from the foregoing to those skilled in the art that changes and modifications may be made without departing from the invention in its broader aspects.

## Claims

1. A method of compiling and accessing subject-specific information from a computer network, the method comprising the steps of traversing links between sites on the computer network, filtering the contents of each site visited to determine relevancy of content, and presenting information on each site deemed relevant for indexing.
2. The method of claim 1, further comprising the step of filtering the contents of a site at least a second time for relevancy.
3. The method of claim 1 or 2, wherein at least one filtering step comprises the steps of presenting the contents to a human editor, approving, by the human editor, if the contents are deemed relevant; and disapproving, by the human editor, if the contents are not deemed relevant.
4. The method of any preceding claim, wherein at least one of said filtering steps comprises the step of passing the contents of the site through a lexicon-based filter, the filter comparing contents of the site with terminology found in the lexicon.
5. The method of claim 4, wherein the step of passing the contents of the site through a lexicon-based filter comprises the step of:
  - (a) comparing the contents of a web page corresponding to the site with the lexicon; or
  - (b) breaking up a web page corresponding to the site contents into component parts and comparing the contents of each component part with the lexicon.
6. The method of claim 5, wherein the step of passing the contents of the site through a lexicon-based filter further comprises the steps of:
  - (a) assigning a weight to the web page based on a result of the step of comparing; and deeming the web page to be relevant if it achieves a high-enough weight; or
  - (b) assigning a weight to each component part based on a result of the step of comparing; and deeming the component part to be relevant if it achieves a high-enough weight.
7. The method of claim 6, wherein the step of assigning a weight comprises the steps of:
  - (a) assigning a weight to each word, term, or expression in the web page that matches a word, term, or expression in the lexicon, according to a weight associated with the word, term, or expression, and accumulating a sum of assigned weights, the sum forming the weight assigned to the web page; or
  - (b) assigning a weight to each word, term, or expression in the component part that matches a word, term, or expression in the lexicon, according to a weight associated with the word, term, or expression, and accumulating a sum of assigned weights, the sum forming the weight assigned to the component part.
8. The method of claim 6 or 7, wherein the step of deeming comprises the steps of:
  - (a) saving the web page and passing it to the step of presenting if it achieves a high-enough weight, and discarding the web page if it does not achieve a high-enough weight; or
  - (b) saving component parts deemed to be relevant and passing them to the presenting step, and discarding component parts deemed not to

- be relevant.
9. The method of any of claims 5-8, wherein the step of passing the contents of the site through a lexicon-based filter further comprises the steps of:
- if at least one component part is deemed to be relevant, passing the web page to the presenting step; and
- if no component part is deemed to be relevant, discarding the web page.
10. The method of any preceding claim, wherein the step of compiling a database comprises the step of, for each relevant site to be stored in the database, assigning a word score to each word appearing on that site.
11. The method of claim 10, wherein the step of assigning word scores comprises the steps of determining all sites found in the database that contain links to the site, and for each word on the site, assigning a word score for that word based at least in part on its presence on each site containing a link to the site.
12. The method of claim 11, wherein the step of assigning a word score for that word further comprises the step of increasing the word score for each site containing a link to the site if the word appears in close proximity to the link.
13. The method of claim 10, wherein the step of assigning word scores comprises the steps of determining all sites found in the database that contain links to the site, and assigning a word score to each word on the site based at least in part on how many sites linking to the site also contain the particular word.
14. The method of claim 13, wherein the step of assigning a word score for that word further comprises the step of increasing the word score for each site containing a link to the site according to the proximity of the word to the link.
15. The method of any preceding claim, further comprising the step of monitoring a depth for each link, the depth being a reflection of relevance.
16. The method of claim 15, wherein the step of monitoring comprises the steps of:
- for a given site being visited, setting depths of any links leading from that site to other sites to a depth of a link traversed to reach the given site;
- if the given site is determined to be relevant in the filtering step, setting the depths of the links leading from that site to zero; and
- If the given site is determined not to be relevant in the filtering step, incrementing the depths of the links leading from that site.
17. The method of claim 16, wherein the step of monitoring further comprises the steps of:
- comparing the incremented depths to a predetermined maximum depth value;
- if the incremented depths exceed the predetermined maximum depth value, discarding the links leading from the given site; and
- if the incremented depths do not exceed the predetermined maximum depth value, traversing one of the links leading from the given site.
18. The method of any of claims 4-17, further comprising the step of replacing the lexicon with a lexicon corresponding to a different subject in order to create a different subject-specific database.
19. The method of any preceding claim, further comprising the step of compiling a database of searchable relevant information.
20. The method of any of claims 10-19, further comprising the steps of: entering a user query, using the user query to search the database, and computing a site ranking for each site associated with information found in said searching step, the site ranking being computed based on said word scores.
21. The method of claim 20, wherein the step of computing a site ranking comprises the step of, for each site associated with information found in said searching step, summing the word scores for that site corresponding to words in the user query.
22. The method of any of claims 1-19, further comprising the steps of permitting a user to enter a query, and searching the database for information according to the query.
23. The method of claim 22, further comprising the step of displaying information found in said step of searching in a hierarchical format.
24. The method of claim 23, further comprising the step of determining a site ranking for each site associated with information found in said searching step, where the determining is according to how interesting at least one of authors and users of the computer network have found the site associated with the information.
25. The method of claim 24, further comprising the step of displaying the results of the user query using the

- site ranking of each item of information found in the search to determine an order in which the results are displayed.
- 26.** The method according to Claim 25, wherein the step of displaying the results of the user query comprises the step of displaying the results of the user query in a hierarchical format according to site ranking. 5
- 27.** A method of ranking the relevance of information stored in a database, the information comprising web pages, the method comprising the steps of: 10
- computing and storing a word ranking for each word, except for stop words, found on each web page; and 15
- in response to a user query, computing a site ranking for each web page found in response to the user query based on the word rankings. 20
- 28.** The method of claim 27, wherein the step of computing a word ranking is performed 25
- (a) according to how interesting at least one of authors and users of a computer network in which each web page is resident have found the web page and/or 30
- (b) comprising the steps of, for each word except stop words on each web page, determining all web pages found in the database that contain links to the web page on which the word appears, and assigning a word score for that word based at least in part on its presence on each web page containing a link to the web page on which that word appears, the word score constituting the word ranking for that word. 35
- 29.** The method of claim 28, wherein the step of assigning a word score for that word further comprises the step of increasing the word score for each web page containing a link to the web page on which that word appears if the word appears in close proximity to the link. 40
- 30.** The method of claim 29, wherein the step of computing a site ranking comprises the steps of: 45
- for each web page found in response to the user query, summing the word rankings for that web page corresponding to words in the user query. 50
- 31.** A computer-readable medium containing software for implementing the method of any preceding claim. 55
- 32.** A system for compiling and accessing information from a computer network, the system comprising a processor, and the computer-readable medium of claim 31.
- 33.** A system that compiles and permits accessing of subject-specific information from a computer network, the system comprising:
- a host computer executing software from a computer-readable medium, the software comprising:
- a smart crawler for traversing the computer network;
- a first filter, filtering out irrelevant sites, and permitting only relevant sites to pass; and
- an indexer indexing the relevant sites; and
- memory, connected to the host computer, for storing indexed subject-specific information.
- 34.** The system of claim 33, wherein the software further comprises at least a second filter.
- 35.** The system of claim 33 or 34, wherein the system further comprises a human-computer interface, and wherein at least one of said filters comprises:
- a presentation of relevant site information received from the smart crawler to a human editor via the human-computer interface; and
- means for receiving input from the human editor, entered via the human-computer interface, as to whether or not to index and store the site in the memory.
- 36.** The system of any of claims 33-35, wherein the first filter or at least one of the first and second filters is lexicon-based.
- 37.** The system of claim 36, wherein the system further comprises an interchangeable computer-readable medium on which is stored the lexicon for the lexicon-based filter, the lexicon containing subject-specific terminology.

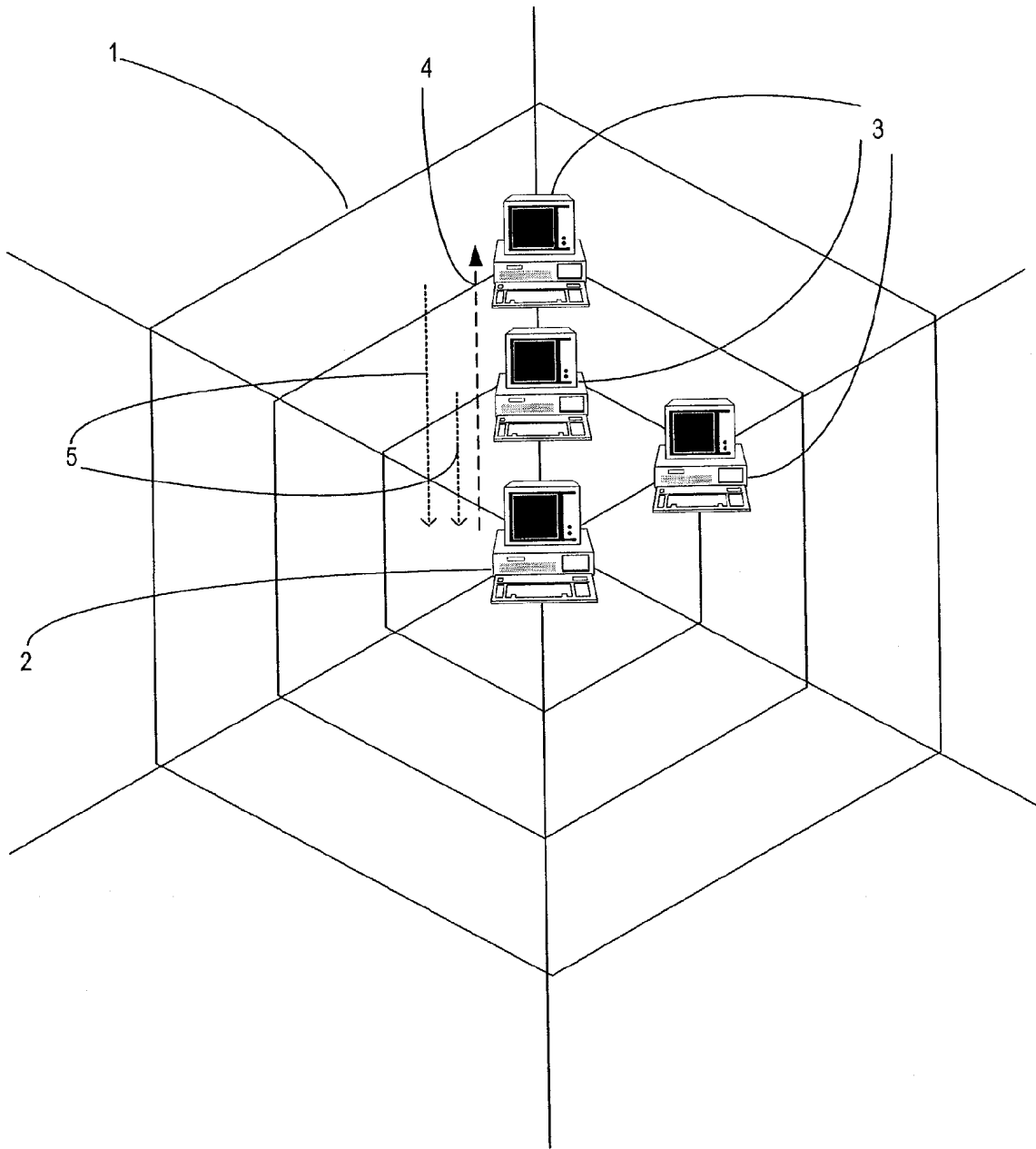


FIGURE 1A

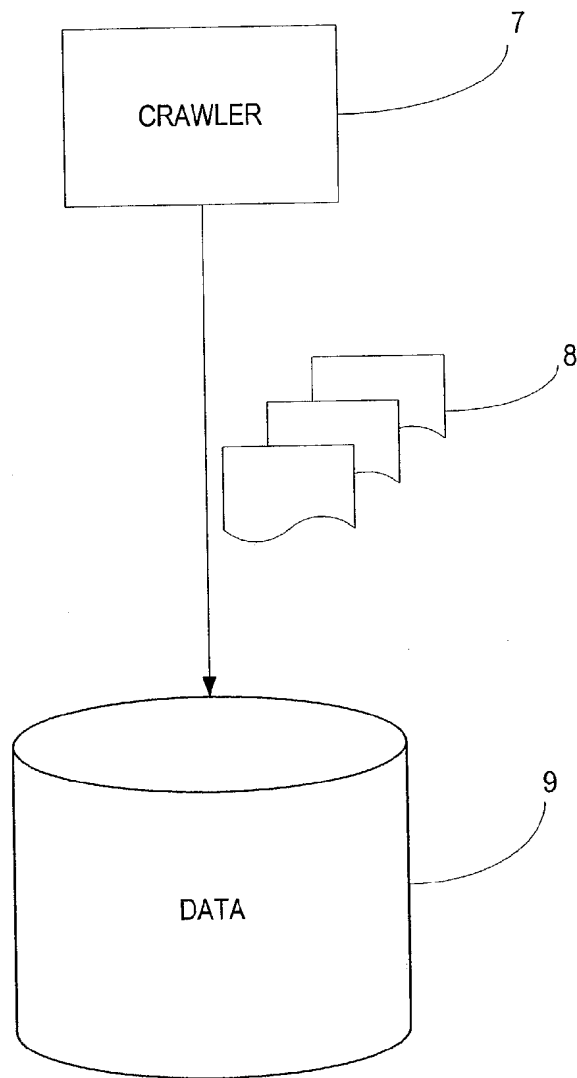


FIGURE 1B  
PRIOR ART

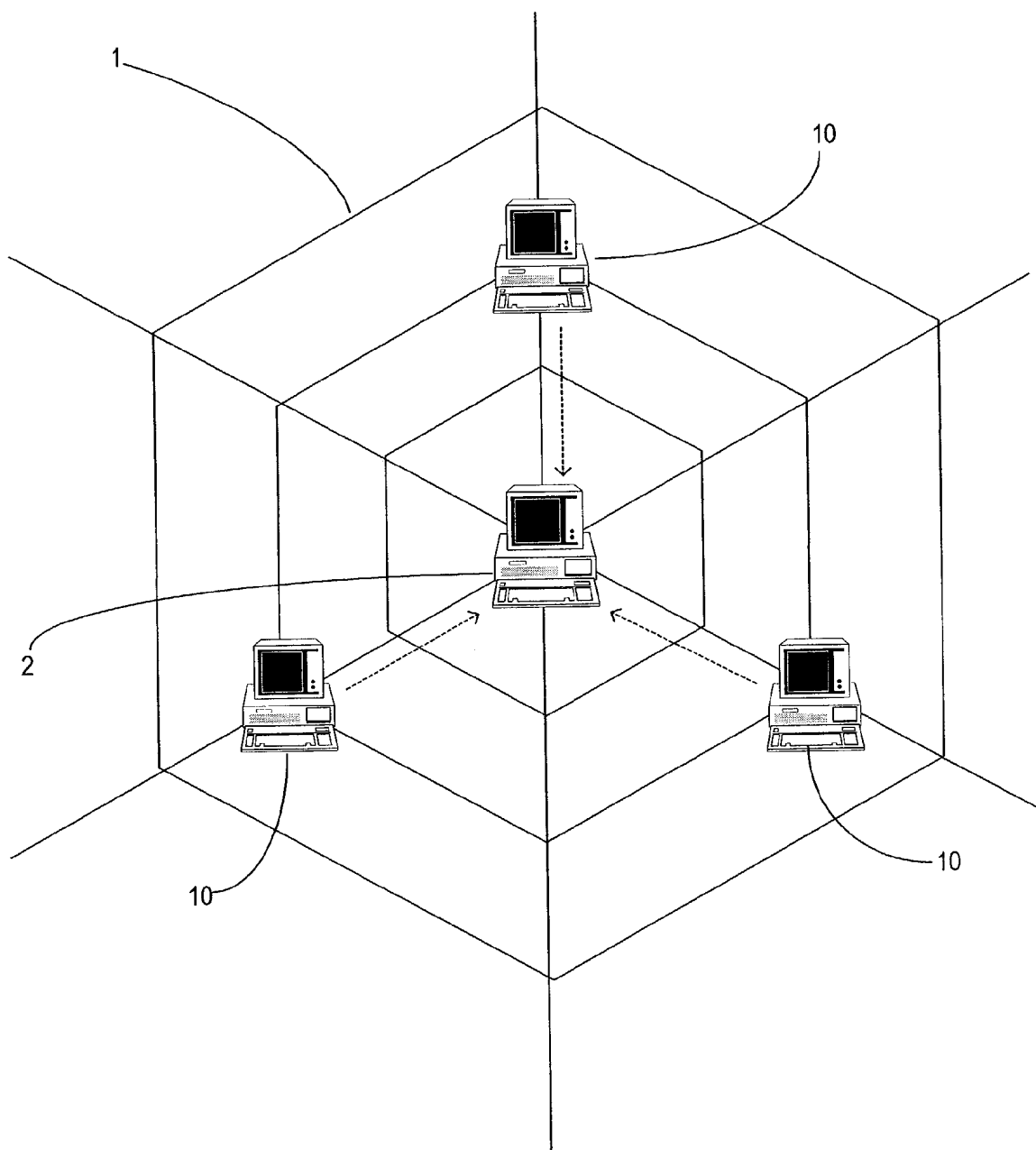


FIGURE 2  
PRIOR ART

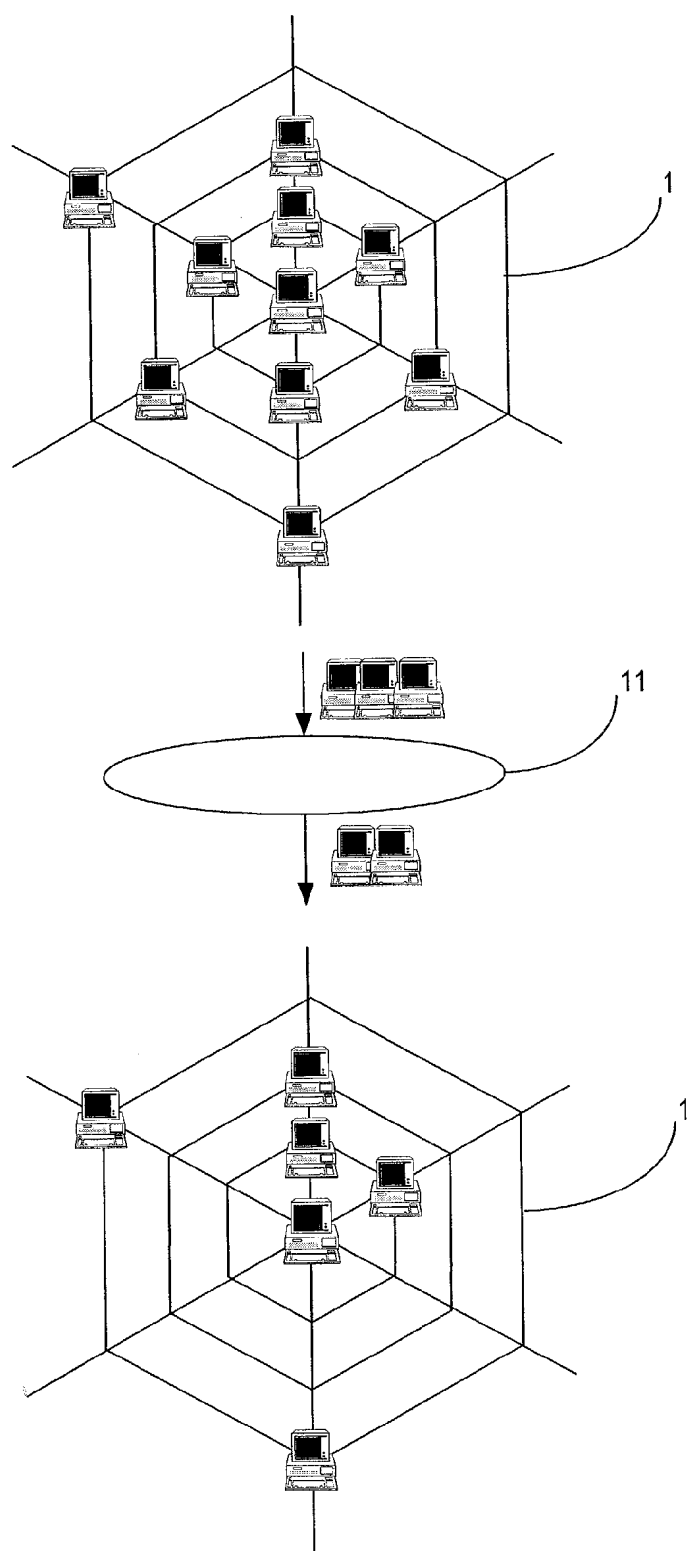


FIGURE 3A

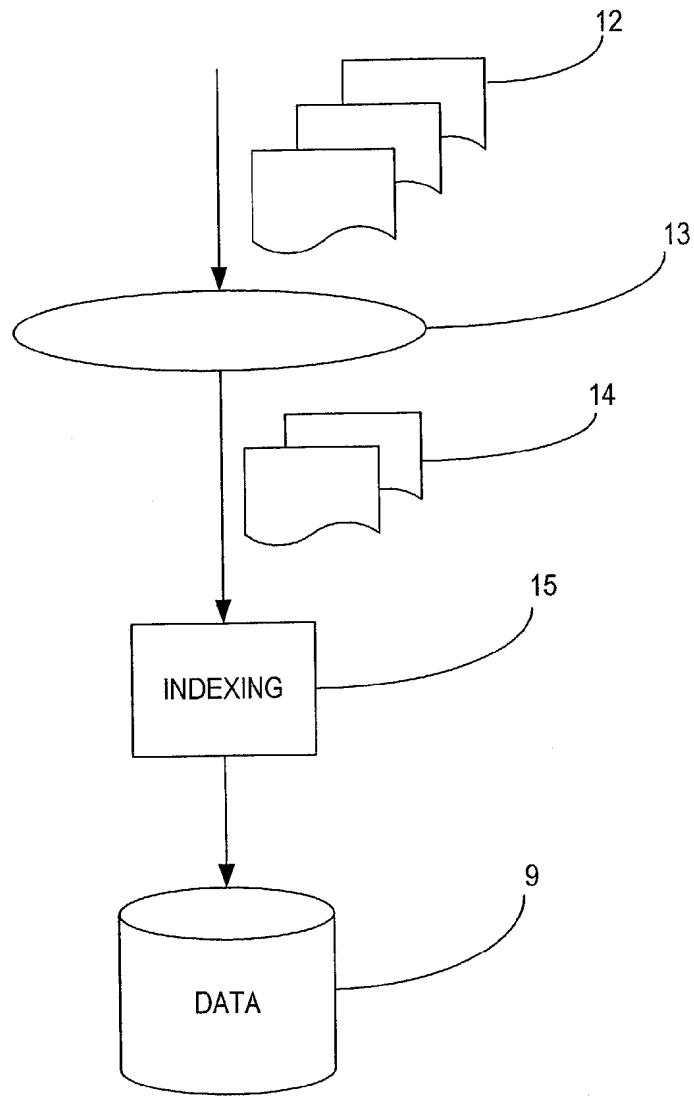


FIGURE 3B

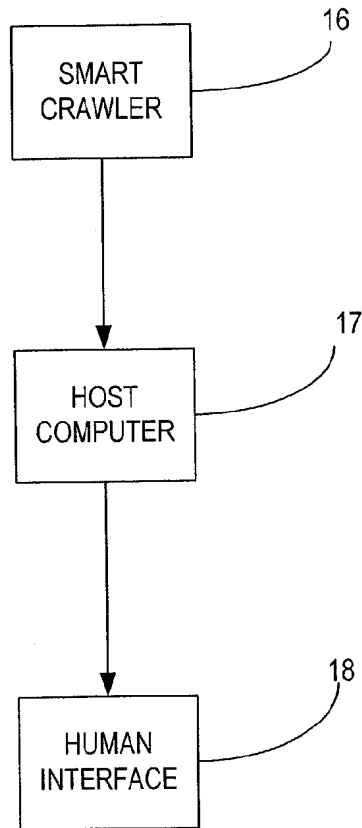


FIGURE 4

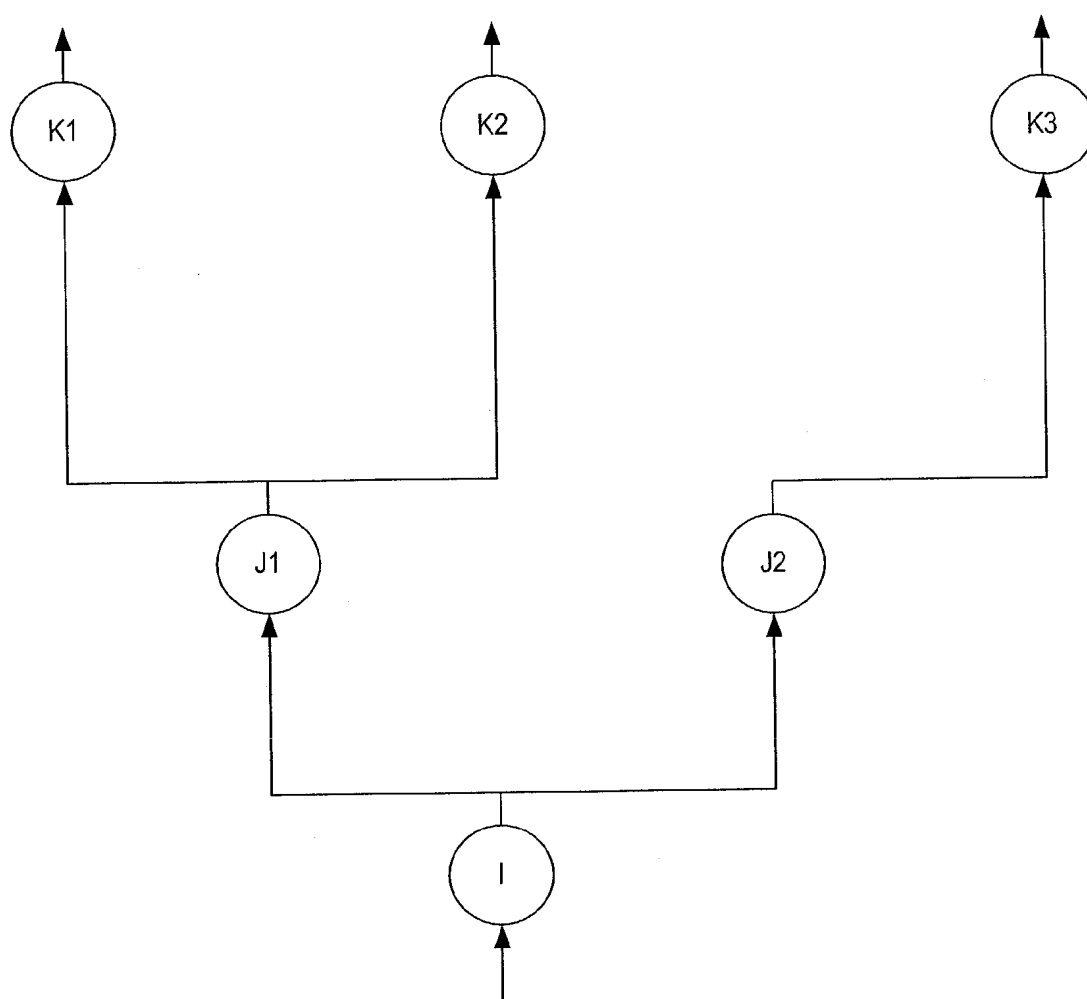


FIGURE 5A

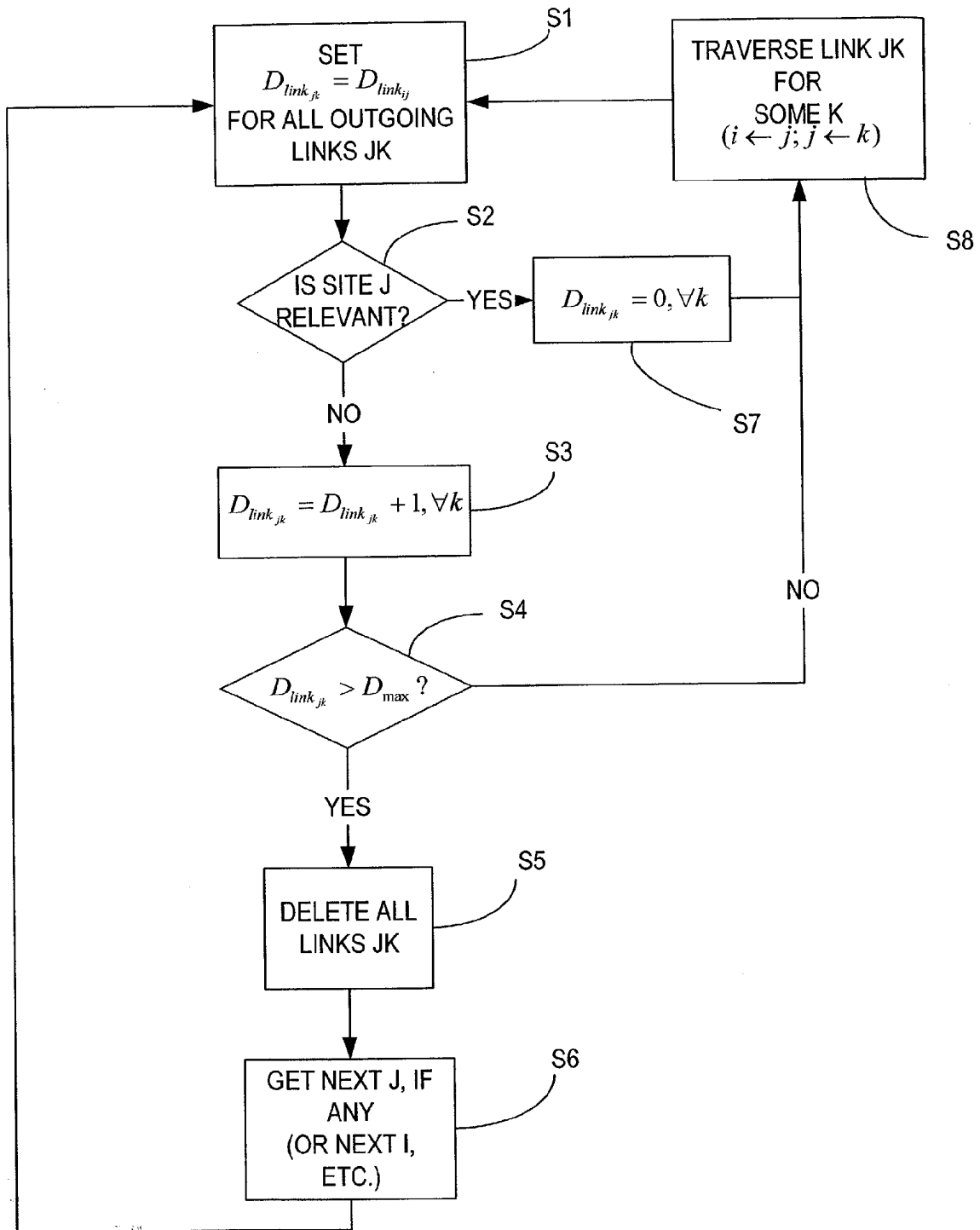


FIGURE 5B

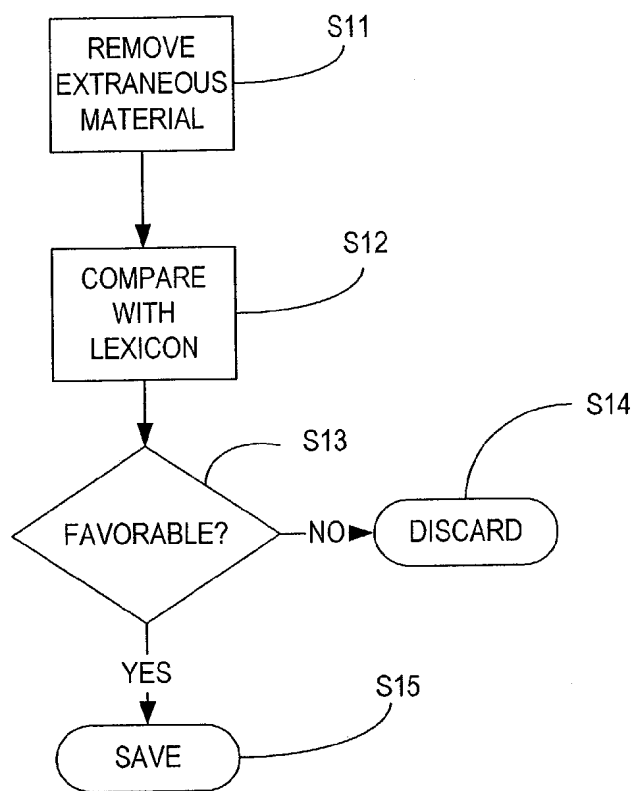


FIGURE 6A

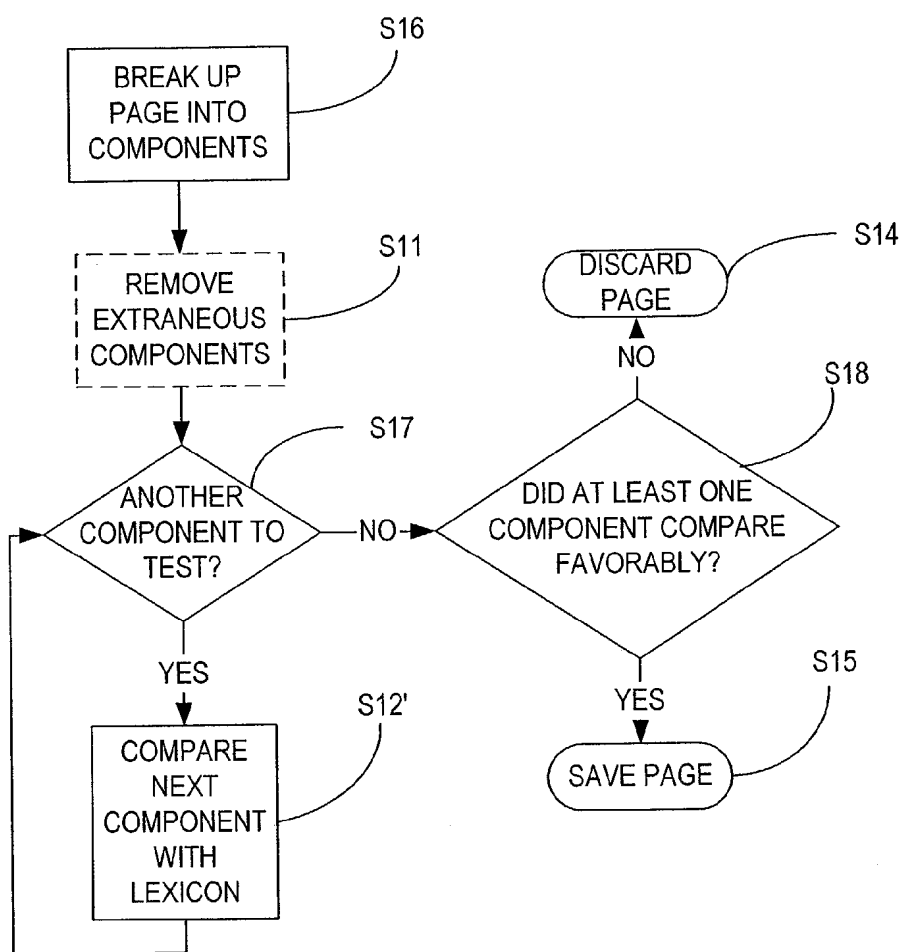


FIGURE 6B

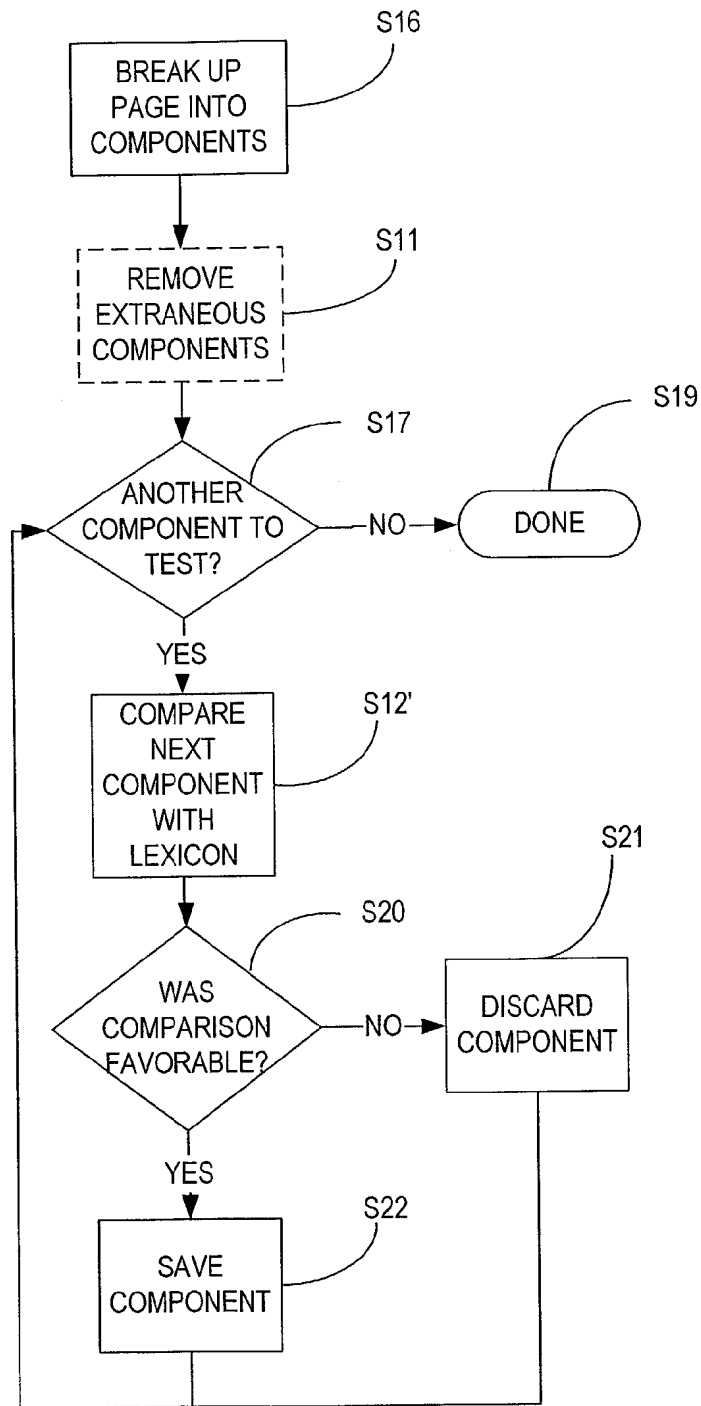


FIGURE 6C

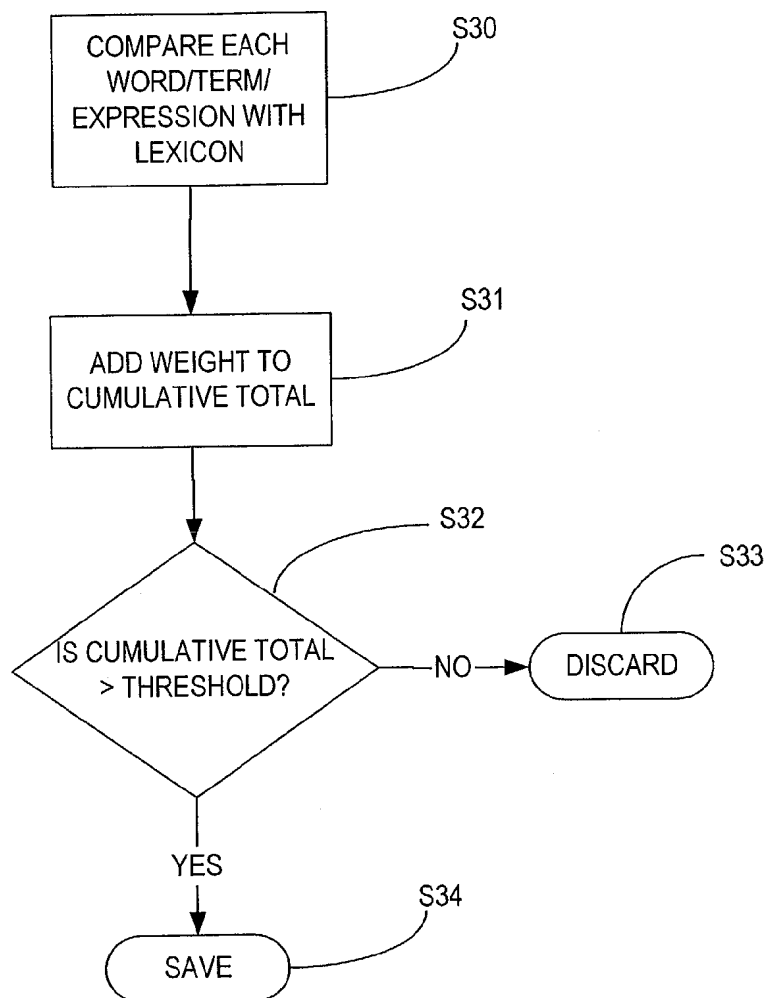


FIGURE 7

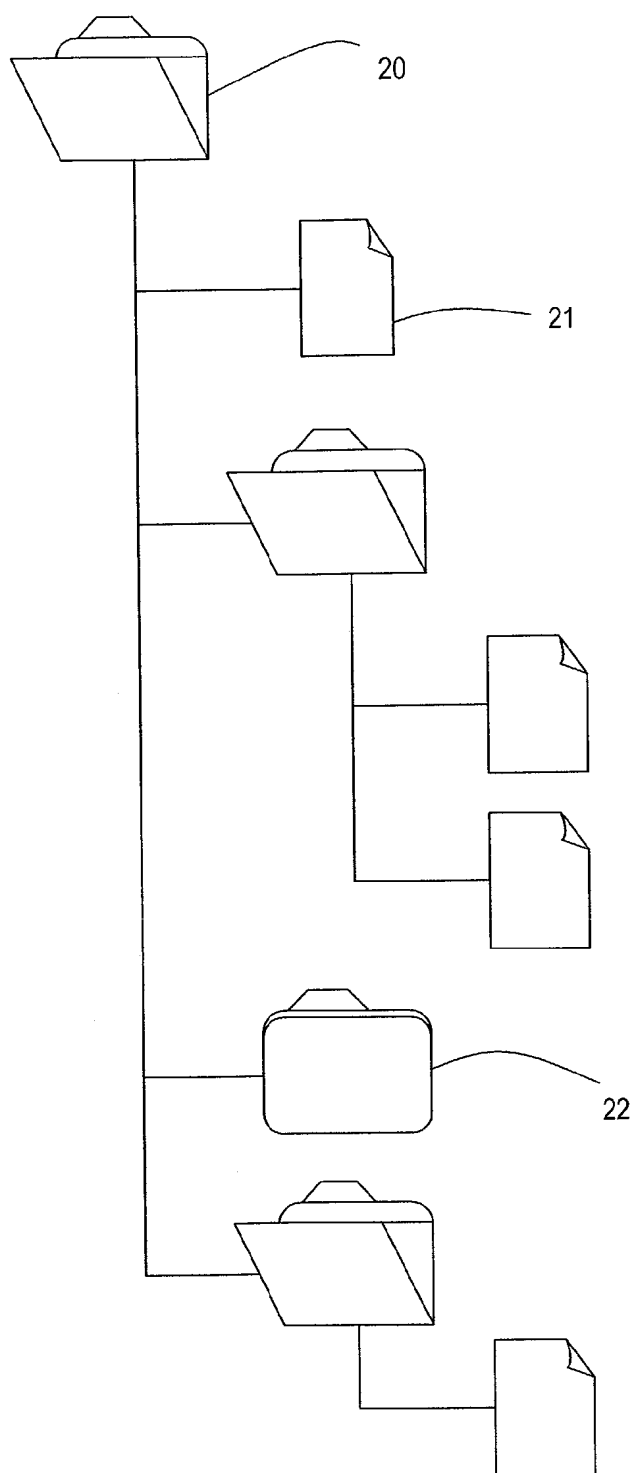


FIGURE 8A

SITE TYPE A	▶	
SITE TYPE B	▶	
SITE TYPE C	▶	
SITE TYPE D	▶	
SITE TYPE E	▶	
SITE TYPE F	▶	
		SITE F1
		SITE F2
		SITE F3

FIGURE 8B